OPEN ACCESS

King's College LONDON

King's Research Portal

Document Version
Early version, also known as pre-print

Link to publication record in King's Research Portal

Citation for published version (APA):
Gold, N. (2014). Trustworthiness and Motivations. In N. Morris, & D. Vines (Eds.), *Capital Failure: Rebuilding trust in financial services* (pp. 129-153). Oxford University Press.

Citing this paper
Please note that where the full-text provided on King's Research Portal is the Author Accepted Manuscript or Post-Print version this may differ from the final Published version. If citing, it is advised that you check and use the publisher's definitive version for pagination, volume/issue, and date of publication details. And where the final published version is provided on the Research Portal, if citing you are again advised to check the publisher's website for any subsequent corrections.

General rights
Copyright and moral rights for the publications made accessible in the Research Portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognize and abide by the legal requirements associated with these rights.

•Users may download and print one copy of any publication from the Research Portal for the purpose of private study or research.
•You may not further distribute the material or use it for any profit-making activity or commercial gain
•You may freely distribute the URL identifying the publication in the Research Portal

Take down policy
If you believe that this document breaches copyright please contact librarypure@kcl.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.

Download date: 05. May. 2023

Trustworthiness and Motivations

1. Introduction: The Principle of Self-Regard

In 1836, James Stewart Mill wrote of political economy, as it was then called, that it presupposes 'an arbitrary definition of man, as a being who inevitably does that by which he may obtain the greatest amount of necessaries, conveniences, and luxuries, with the smallest quantity of labour and physical self-denial with which they can be obtained' (Mill, 1836, V.46). Despite the rise of behavioural economics, this is still the standard picture. As a widely-used graduate textbook in microeconomic theory states: 'A defining feature of microeconomic theory is that it aims to model economic activity as an interaction of individual economic agents pursuing their private interests' (Mas-Colell, Whinston, & Green, 1995). Standard models assume not only that people are *self-interested,* in the sense of being concerned with their own well-being, they are also assumed to be *selfish,* in the sense of *only* being concerned with their own well-being, and even *self-regarding*, in that their well-being merely concerns themselves and does not reference any other agent—a kind of solipsism or 'unsympathetic isolation'.[1] Hence I will call this assumption about motivation the *principle of self-regard.*[2] It is often traced back to Adam Smith's *Wealth of Nations*, where he famously wrote, 'It is not from the benevolence of the butcher, the brewer, or the baker that we expect our dinner, but from their regard to their own interest. We address ourselves, not to their humanity, but to their self-love, and never talk to them of our own necessities, but of their advantages.' (Smith, 1776, Ch.2.)

The principle of self regard became increasingly important in the late 19th century, with the work of economists such as Alfred Marshall and Francis Edgeworth, whose analyses emphasized the way in which the interaction of individual agents causes economic outcomes. They pioneered a mathematical model of behaviour in which individuals maximize utility and firms maximize profits, subject to constraints on their budgets and resources. This is the core of neoclassical

[1] The term 'unsympathetc isolation' is due to Edgeworth (1881, p.12).

[2] This definition of self-regard follows from Mill's definition of self-regarding conduct in *On Liberty* (I.9), as that which 'merely concerns oneself'. Mill's concern was with actions, and whether they would have any harmful effect on others, but his adjective could just as well be applied to people's interests.

economics, which is in the current mainstream of the subject.[3] In terms of the underlying mathematics, 'utility' is an empty placeholder which includes anything that might make an agent choose one option over another. In other words, the theory of behaviour that the model represents is one where individuals pursue their interests, where 'interests' can be understood in the loosest possible sense of the term, as anything that a person would like to achieve.[4] However, in practice, the content of 'utility' needs to be specified if models are to have any predictive or descriptive power—which is equivalent to delimiting a person's interests. When interpreting and applying economic models, utility is usually taken to be a function of the agent's own consumption of goods and services, and agents to be self-regarding. Agents are, as Edgeworth put it, in a state of 'unsympathetic isolation' (Edgeworth, 1881, p.12).

The principle of self-regard was not supposed to be taken as a theory of human nature. Although it is often attributed to Adam Smith, he most certainly did not endorse it. Despite the above, often quoted, passage from the *Wealth of Nations*, Smith opened his earlier *Theory of Moral Sentiments* with a contradictory empirical claim, 'Howsoever man may be supposed, there

[3] Specifically, it is the mainstream way of modeling individuals and firms. But not all mainstream economics papers model individuals or firms.

[4] The 'utility' terminology can cause confusion because of its etymology. It was introduced into economics alongside formal methods, in the so-called 'marginalist revolution', which showed how prices depend on the value of the last (or marginal) unit consumed or produced. In the influential work of William Jevons, marginalism was associated with a theory of value that was based on Bethamite utilitarianism and the idea that there is a monistic and measurable pleasure-pain index: hence 'marginal utility' (Jevons, 1871). However, as Bentham's theory fell out of favour, economists also distanced themselves from it, culminating in the 'ordinal revolution' of the 1930s, when economists rejected the idea that there was a cardinal scale of utility. Utility functions were still used but they only represented an ordinal valuation, i.e. consumers were assumed to be able to rank commodity bundles but not to be able to quantify these judgments. The fact that these ordinal preferences could be represented with a mathematical function meant that, in the formalism of the model, consumers are represented as 'maximizing utility'. However, despite the continued use of the word 'utility', economics had been de-coupled from utilitarianism: the primitive concept was that of a preference ranking, and no assumptions were made about what considerations under-pinned the ranking. To maximize utility is just to choose the most preferred consumption bundle from those available.

are evidently some principles in his nature, which interest him in the fortune of others, and render their happiness necessary to him, though he derives nothing from it except the pleasure of seeing it.' (Smith, 1759). For Smith, humans are fundamentally social beings and have other-regarding motivations. Furthermore, the *Wealth of Nations* builds on the discussion that was in the *Theory of Moral Sentiments*, where it is made clear that self-interest is set within the wider context of social obligations and that it is associated with the classical virtue of prudence, not the vice of greed (Smith, 2013). It has become commonplace to take passages from the *Wealth of Nations* out of context and to forget that Smith's writings pre-date the neoclassical idea of individuals as utility maximizers.

Even the fathers of neoclassical economics merely took the principle of self-regard to be a good approximation of motivation in certain domains. Mill acknowledged that conduct could depend on 'the feelings called forth in a human being by other individual human or intelligent beings, as such; namely, the *affections,* the *conscience,* or feeling of duty, and the love of *approbation.*' (Mill, 1934, V.34, italics in the original.) However, he considered these motivations to be the subject matter of philosophy. Edgeworth thought that people care about the welfare of others, even indicating how concern for others could be incorporated into his mathematical framework, but he believed that the principle of self-regard was a reasonable assumption in both war and trade (Collard, 2001).

Reticence about the domain of economic theory was also imposed by the need for cardinal measurement of utility in early formulations of the neoclassical model. It was more plausible that people could make the required numerical comparisons if they only had to compare material satisfactions. However, once the original cardinal foundations (and the association with Utilitarianism) were rejected in favour of the modern ordinal ones, which only require a ranking of outcomes, that paved the way for the utility maximizing model to become an all-encompassing theory of human behaviour (Lewin 1996; Mandler 2001). But, if the ordinal framework is augmented with the principle of self-regard, then the expansion of the economic approach looks more dubious.

In the policy domain, the popularity of the principle of self-regard may owe something to Smith's idea that a person who intends only his own gain is led by an 'invisible hand' to pursue the good of society, 'more effectually than when he really intends to promote it' (Smith, 1776, Ch.2.). Not only has the principle of self-regard seemed like a reasonable assumption but, if people would act according to it, then it would promote good outcomes. However, the principle of self-regard narrows the range of tools that are available for policy makers. A consequence of using Mills'

'arbitrary definition' is that interventions are limited to financial incentive schemes, or regulations that are enforced by the threat of fines or prison. One argument of this chapter is that excessive focus on the principle of self-regard obscures other–potentially more effective–policy interventions.

Even within the domain of political economy, both Marshall and Mill explicitly recognized that the principle of self-regard was a simplification. Marshall thought that the focus on self-regard was justified because 'the steadiest motive to ordinary business work is the desire for the pay which is the material reward of work' (Marshall, 1890, Book 1, Ch.2. V1.). However, he immediately followed this with the statement that, 'Everyone who is worth anything carries his higher nature with him into business'. Similarly, Mill conceded that the principle of self-regard was a simplification, 'treating the main and acknowledged end [of behaviour] as if it were the sole end' (Mill, 1934, V.38) This led Mill to conclude that the resulting 'approximation' of behaviour might need to be corrected to take account of other impulses (Mill, 1934, V.34). To the extent that we are influenced by other motives, models based on the principle of self-regard will fail to explain or predict events, and policies based on the principle of self-regard will not have the desired effects.

In this chapter, I will argue that we need to move beyond self-regard when we formulate regulations for finance. Self-regard is not a good assumption and behaviour that is based on it will not produce good outcomes. That is because the principle of self-regard precludes an important sort of trust and trustworthiness, which are based on non-self-regarding motivations, and which we rely on in finance. This implies that, when formulating policy, we should consider how to design institutions and regulations so that they induce the relevant non-self-regarding motivations, and that we should design financial institutions so that they attract employees who are more likely to have those motivations.

2. Beyond Self-Regard: A richer account of motivations

In economics, it is standard to assume that an agent is only motivated by her own material rewards and punishments, and to investigate the optimal way to structure incentives given these self-regarding motivations. There are two cross-cutting objections to this project: that people are not only motivated by *their own rewards*, and that people are not only motivated by their *material rewards*. My focus is on the first of these objections but, before I explore some of the ways in which people may be non-self-regarding but, before I do that, it is worth explaining how the two objections relate to each other.

When a person acts in order to get an 'apparent reward' (or avoid a punishment),

psychologists say that she has an *extrinsic motivation* (Deci, 1975). This is in contrast to *intrinsic motivation*, which does not involve apparent external rewards. Examples of intrinsic motivation include completing a task because it is fun or because one is obliged, as opposed to doing a task because one will be paid or punished depending on completion. We might think of extrinsically motivated behaviour as that which aims to get an external reward, supplied by some other agent.

The distinction between intrinsic and extrinsic motivations divides up behaviour differently from that between self-regarding and non-self-regarding behaviour. For instance, Marshall acknowledged that, in the economic domain, as well as being motivated by the desire for pay some people are motivated by a desire for approbation or by the pleasure of doing skilful work (Marshall,1890). The pleasure of doing skilful work is clearly an intrinsic motivation, but it is also a self-regarding one. The desire for approbation is arguably an extrinsic motivation, but it is certainly other-regarding, as it challenges the assumption of unsympathetic isolation.

My concern is with the self-regard of neoclassical agents: their goals never depend on other people, and their behaviour is always about the achievement of those goals, never about how they act in the pursuit. Hence neoclassical economics neglects some of the goals that we pursue which, whilst selfish, essentially depend on other people. It also entirely disregards two important classes of motivation, prosocial and procedural, where the goals pursued are not the narrow (selfish) self-interest of the agent.[5]

Prosocial motivations

People are not only concerned with *their own* outcomes. They may be concerned with the outcomes of others. (In the economic model: agents' utility may be a function of others' outcomes as well as their own.) The desire to improve the outcome of others is a *prosocial* motivation; its opposite, which is rarely studied, could be considered an *antisocial* motivation, the desire to diminish the outcomes of others.[6] These are types of other-regarding motivations.

[5] Note that my concern is with proximate goals. People may get a 'warm glow' from non-self-regarding behaviours, so that pursing such goals is still, in a sense, self-interested or welfare-enhancing. I do not have a stake in the debate about whether such self-interest is always the ultimate goal of action. For more on that debate, especially arguments and evidence that helping behaviours are not always the result of an ultimately self-interested motivation, see Sober and Wilson (1998), Batson and Shaw (1991), and Batson (2011).

[6] For some exceptions, see Zizzo and Oswald (2001), Abbink and Sadrieh (2009).

Prosocial motivation covers a variety of ways in which we may be concerned about the outcomes of others. The one that has attracted most attention from researchers is *altruism*, the concern for the outcomes of another or others (e.g., Collard, 1978; Fehr and Fischbacher, 2003). Another prosocial motivation that is increasingly attracting attention is the concern for the outcome of one's group (e.g. Bacharach, 2006; Sugden, 1993). This differs from altruism because it stems from a common category membership and is a concern for the collective outcome of 'our' or 'my' group, whereas altruism is the inter-personal promotion of 'your' or 'their' welfare (Brewer & Gardner, 1996). The outcome of the group need not necessarily reduce to the outcomes of the individual members, nor must improvements in the group outcome reflect improvements in the outcomes of individual members, although we might think that it is likely to do so, or that it will do so in a well-functioning group (Gold, 2012). Experimental manipulations that increase group identity lead to more prosocial behaviour (e.g. Brewer and Kramer, 1986). It is difficult to disentangle whether this is caused by concern for the outcome of the group or by altruism because, particularly in small groups, increasing group identity may also increase inter-personal altruism between members of the group. However, it is important to distinguish the two concerns conceptually because they may lead to different outcomes (Bacharach, 1999; Gold & Sugden, 2007).

Procedure-regarding motivations

People are not only concerned with their own *outcomes*. The aims of human behaviour are not always focussed on end states. For example, people may want to behave fairly, to behave morally, to follow norms, or to abide by standards (e.g. professionalism, doing a good job according to standards in one's field). They are not so much interested in the outcome per se as in the way in which it is achieved or the principle on which they act. These motivations are procedure-regarding.

It may be possible to interpret some cases of procedure-regard as the achievement of outcomes or end states (for instance, we might think of fairness as being about achieving equal outcomes, or morality—on consequentialist views—as being about implementing the best outcomes), or to fit them into a modelling framework of means-end reasoning.[7] However, even if that is possible, it will get the order of explanation wrong. Sometimes people desire to follow a

[7] For the possibilities and limits of this modelling approach, see, e.g., Broome (1992), Brown (2011).

procedure, often a normative rule, for its own sake, not as a means to an end; the ensuing outcome is secondary to the choice of procedure.

Selfish yet other-regarding motivations

Even allowing that agents are completely selfish and concerned with their own outcomes, these outcomes may reference other people. For instance, agents may care about being esteemed by others, and pursuit of esteem may motivate their behaviour (Brennan & Pettit, 2004; Offer, 1997 and in this volume). Here we have a sense in which agents may be other-regarding despite being completely selfish: they are concerned with the opinions of others.

We can use esteem and regard to incentivise behaviour; they increase our repertoire of extrinsic rewards. Even if undertaken for purely selfish reasons, the pursuit of esteem and regard may indirectly lead people to care about the outcomes of others—because how others perceive your intentions and your contributions to their outcomes will affect the attitudes they hold towards you.[8] In George Elliot's *Mill on the Floss* (Book 1, ch 6), Tom and Maggie Tulliver split a jam puff. Tom does his best to divide it equally but fails, and Maggie urges Tom to take the best bit. However, Elliott remarks 'I fear she cared less that Tom should enjoy the utmost possible amount of puff, than that he should be pleased with her for giving him the best bit', establishing that Maggie is not an unselfish character, despite her other-regarding behaviour.

We are not purely self-regarding creatures. As well as copious experimental evidence, we know this from introspection and by observing everyday life. Furthermore, as I will go on to argue, non-self-regarding motivations are essential to a proper understanding of trustworthy behaviour.

3. Trust, Trustworthiness, and Motivations

Trusting is a risky business. A truster makes herself vulnerable to her trustee, exposing herself to the risk that her trust will not be fulfilled. We can think of trust as a three place relation: *A* trusts *B* to *X. A* is the truster, *B* the trustee, and *X* is an undertaking with an outcome that *A* cares about. A person (or institution) who fulfills the trust that is placed in them is trustworthy.

In economics, this minimal definition is taken as sufficient: trust and trustworthiness are defined as behaviours (see, e.g., Fehr, 2009) and the neoclassical way of studying trust is to ask to

[8] See Rabin (1993) for a classic economic model that includes intentions, and Falk, Fehr and Fischbacher (2008) for recent evidence that intentions matter.

how trustworthy behaviour can be sustained, based on self-regarding motivations. According to this approach, trustworthiness can be ensured by threatening punishments and offering rewards, structuring the trustees incentives so that it is in her self-regarding interest to be trustworthy. One type of reward is the expected benefit from future encounters so, if the truster and the trustee have a continuing relationship, then there is an incentive for a self-regarding agent to be trustworthy (Hardin, 1994; 2004).

However, the idea that trustworthiness is based on self-regard is empirically and theoretically inadequate. Empirically, there is plenty of evidence that people behave in a trustworthy manner, even when that leaves them worse off. (See discussion of the trust game, below). Theoretically, most philosophers reject the neoclassical analysis of trust. Philosophical analyses also use the behavioural definition of trust but they take it as a starting point: a necessary condition that must be further augmented because, as it stands, the neoclassical analysis conflates trustworthiness with reliability.

The concern to distinguish trustworthiness from reliability stems partly from the fact that, unlike reliability, trustworthiness is often considered to be a moral virtue. Reliability is a property that may be possessed by mechanical objects. For example, we may rely on our alarm clock to wake us up in the morning, but we do not trust it. However, reliability is not just a property of mechanical objects, it can also apply to human agents performing intentional actions. For instance, the philosopher Kant was in the habit of taking a walk at 3.30pm every day and he was so punctual that his neighbours could set their clocks by him. Imagine a neighbour who used Kant's walk to time the school run, almost as though he were an alarm clock. She would have been relying on him in order that her children would not be left waiting by the school gates, but it would be wrong to say that she trusted him. Furthermore, if Kant had been late for his walk, causing her to be late for the children, then he would not have been culpable. But someone who breaches a trust is prima facie culpable. This difference between reliability and trustworthiness is reflected in the moral psychology of trust. When someone we trust lets us down, we feel betrayed; but disappointment is the appropriate attitude to being let down by someone or something we rely on. There is a normative element to trust, which is not present in reliance.

It is tempting—but wrong—to conclude from the above examples that the difference between trustworthiness and reliability relates to 'intending to X'. After all, a car does not intend to start and, although Kant did something intentionally, what he intended was to 'take a walk at 3.30pm' and not 'ensure that his neighbor be on time for the school run'. The neighbour is simply relying on Kant's predictable punctuality, in much the same way that she would depend on a

predictable alarm clock. So it seems that 'intending to X' is necessary for trustworthiness. However, in most of the philosophical literature it is not sufficient that the trustee intends to X, she must also do X for the right sort of reason.

There is widespread agreement amongst philosophers that behaviour that is motivated by the pursuit of rewards or the avoidance of punishment is reliable rather than trustworthy. As Annette Baier puts it, 'We may rely on our fellows' fear of the newly appointed security guards in shops to deter them from injecting poison into the food on the shelves, *once we have ceased to trust them*.' (Baier 1986, p.234, my italics). This stipulation relates to issues of moral psychology and culpability. If we use an incentive system to motivate someone and it fails to work then the fault lies with our design of the system and we should feel disappointed, not betrayed. Hence the general agreement that trustworthy behavior must be sufficiently 'internally driven' (Holton, 1994, p.66).

Philosophers disagree about what the intrinsic motivation involved in trustworthiness must be. A popular account is given by Annette Baier (1986), who argues that trustworthy behaviour is motivated by 'good will', a motivation to take care of something the trustee cares about. Baier identifies trustworthy behaviour as driven by a concern for the truster's outcomes, i.e. it is a type of prosocial behaviour. Karen Jones (1996) argues that goodwill is not enough and that, in addition (in order to exclude cases where someone is reliably benevolent), the trustworthy person must be directly and favourably moved by the fact that someone is counting on her.

In contrast, a compelling recent account of trust, due to Katharine Hawley (2012), locates the difference between trust and reliance in terms of 'commitments'.[9] According to Hawley, trusting someone is a matter of relying on them to meet their commitments, and trustworthiness is a matter of adjusting one's behaviour to one's commitments. Hence the virtue involved in

[9] Note that this is a different usage from the one that is most commonly used in economics. In philosophy, 'commitment' is a normative term and, for Hawley, it is supposed to connote something similar to an obligation (although commitments and obligations are different and she argues that it is commitment, and not obligation, that is important for trustworthiness). This normative usage differs from the use of 'commitment' which is often found in economics, whereby 'commitment' is a descriptive term connoting a form of binding, in contrast to 'discretion' or 'flexibility'. It also differs from the well-known use of the word by Amartya Sen (1977), where 'commitment' is an attitude that transcends self-interest and may result in a choice that does not maximise the agent's welfare.

trustworthiness is that of living up to one's commitments. It is obvious that we can acquire commitments through explicit promises and contracts, but commitments may also arise without an explicit agreement, for example through accepting a role or via social conventions. Hawley says that to ascribe trustworthiness in some specific instance, it is enough for the trustee to behave in accordance with her commitment. She need not be motivated by the commitment. However, when we talk of the virtue of trustworthiness, we indicate a general trustworthiness, someone who will fulfil whatever commitments they have. So we might expect that, 'a generally trustworthy person will often meet her commitments simply because they are her commitments' (Hawley, 2012, p.16). We can classify this as a procedure-regarding motivation, for fulfilling commitments.

Unlike the self-regarding account of trust used in economics, most philosophers favour an account that involves non-self-regarding motivations. However, there is disagreement about how best to formulate a motivational account of trustworthiness; whether it involves a species of prosocial motivation or whether the motivation is ultimately procedure-regarding. As a philosophical theory, I favour the procedural account. But much of what I will go on to argue is compatible with either account because the fiduciary duties that exist in finance involve the commitment to promote the client's interests. What the goodwill and the commitment accounts have in common is that trustworthy (as opposed to reliable) behaviour stems from non-self-regarding, and non-selfish, motivations.

4. Strong Trust

The behavioural definition of trustworthiness used in economics will label some actions as 'trustworthy' that an account of trustworthiness that also stipulates motivations would label 'reliable'. Rather than get into a disciplinary dispute about terminology, we can grant a behavioural use of the word trust (and trustworthiness) but distinguish between 'strong trust' and 'weak trust' (and strong and weak trustworthiness), where *strong trustworthiness* must include a non-self-regarding motivation on the part of the trustee, but *weak trustworthiness* can be the result of any motivation. Hence, weak trust will include cases that some philosophers would claim are only instances of reliance and are not 'really' trust.

Even my definition of strong trust will fail to satisfy most philosophers because it incorporates all non-self-regarding motivations (including selfish but other-regarding motives, such as the pursuit of esteem). I do not need to be invested in the dispute about which non-self-regarding motivation is the 'correct' one because I am concerned with the psychology of

trustworthy behaviour and the motivations that are its wellspring, not the sources of its normativity. So, for instance, the fact that people have a 'sense of moral duty' is relevant, but whether people really have such duties or whether their moral psychology reliably tracks them are of no import to my argument. My argument for using strong trust is consequentialist, that it gets good results. I am sure that sometimes we have a moral duty to be trustworthy. But the foundation of morality is a thorny issue; philosophers don't even agree about why trustworthiness is a virtue. So it is a good thing that we can construct an argument for strong trust in finance without straying into the terrain of moral duties.

We need strong trust. It is more efficient than weak, and we rely on it in most everyday transactions. It has a priority over weak trust, as I will explain below. I will also argue that, in at least some financial transactions, strong trust is important because weak trust may not be a possibility.

The existence of strong trustworthiness

We can be sure that strong trustworthiness exists: experimental economics gives us the evidence. The 'trust game' has two players, a 'sender' and a 'receiver'. The sender is endowed with $10. She may choose to transfer some or all of it to the receiver. Any money that is transferred is multiplied up by the experimenter. Then the receiver gets to choose whether to send any of that money back. By sending money, the sender opens up the possibility of mutual gain because the experimenter increases the pot of money. But, by sending, she also exposes herself to risk because there is no guarantee that the receiver will send any of that money back. If the sender sends money and her trust is not fulfilled, then she is worse off than if she had sent nothing.

If both agents in a trust game are rational and self regarding, then the receiver will never transfer any money back, the sender can predict that there would be no back-transfer if she sent money, and hence she sends none. The sender correctly anticipates that any trust would be breached, so she never places any trust.

However, people trust and are trustworthy. In the original trust experiment conducted by Berg, Dickhaut and McCabe (1995), virtually all senders sent at least some money, the average amount sent was over $5, and roughly one third of the receiver reciprocated by sending back more than was originally sent. The game was only played once, so there was no tangible benefit of sending back money. Therefore those who reciprocated were strongly trustworthy. This is even

more noteworthy because the experiment was 'double blind', i.e. participants were anonymous to the experimenters as well as to each other.

The efficiency of strong trust

Strong trust is both more demanding than weak trust and more stable: it is robust to a wider range of counter-factual situations. Trustworthy behaviour that is only compelled by a desire to avoid punishment will no longer be trustworthy in the absence of that punishment. Robustness is an important property. For example, imagine two parties trying to broker an agreement for mutual benefit. If they can only rely on each other's weak trustworthiness, then they need to construct a contract containing clauses to cover every possible eventuality. If they are motivated by strong trust, then an incomplete contract may suffice. Since most contracts are by necessity incomplete, in the absence of strong trust contractors would often have to fall back on costly legal processes. It is clear from this that, where strong trust is available, it is more efficient than weak.

The priority of strong trust

Strong trust is ubiquitous. It is involved in virtually all market transactions. When we hand over our money to the butcher, the brewer, or the baker, we expect that they will hand over the goods in exchange. According to the principle of self-regard, this behaviour must be motivated by reputation—if it is known that a tradesman does not hand over the goods then people will not trade with him in the future—or by the threat of recourse to legal action. But these are not what really motivates people to complete transactions. When we hand over our money to the butcher, the brewer, or the baker, we trust that they will hand over the goods in exchange, and that is based on the expectation of strong trustworthiness.

In case you are in doubt about this, consider another example, that of taking a taxi. Payments for services always have an element of asymmetry, with the payment point being either before or after the service (and the situation where payment is split between these two points in time is relatively rare). The party who performs their part of the bargain first must trust the other party to complete. When hailing a cab on the street, it is normal to pay for the ride after being dropped off at one's destination and it is unlikely that the rider will ever be picked up by the same taxi driver again, or could be identified and excluded by the community of taxi drivers. But it is

extremely rare for the rider to run off at the end without paying.[10] I have never even considered doing that; nor, I suspect, have most people.

This brings us to the relation between trust and laws, and the priority of strong trust. When we receive goods and services, a legal framework for enforcement of payment does exist and, in extremis, we could have recourse to the law. But the neoclassical idea that we need the law in order to trust (in a weak sense) gets things the wrong way round. Effective laws codify behavioural standards that (most) people are already motivated to live up to and, when there is widespread disagreement with the standard, then the law is not enforced—think of laws against cannabis, homosexual intercourse, tax evasion in some Mediterranean countries. A consequence of this is that legal changes alone, without corresponding changes in social norms, have limited efficacy as a mechanism for behaviour change, as has been found by those who seek to eradicate female genital cutting (Mackie, unpublished). Only when the majority of people are already prepared to comply with the law can the authorities enforce it on a minority of deviants. In other words, it is only because the majority of people are trustworthy in a strong sense that the law can be used to ensure that everybody is trustworthy in the weak sense.

5. Strong Trust in Finance

Strong trust is a feature of all transactions, but it is particularly important in finance. Financial products are often very complicated, there is a whole chain of transactions between the initial seller and the end user, and the end user typically is not in a good position to assess the product. There is *asymmetric information*, where the seller knows important facts about the product that are unknown to the buyer. Asymmetric information is not unique to finance, and it has been much studied by economists. The standard example of asymmetric information is the *market for lemons*, a model of the used car market (Akerlof, 1970). The amount that a buyer is willing to pay depends on the quality of the car she is being offered, but she does not know whether the car is good quality or whether it is a 'lemon'. If she buys a car then she bears a risk of over-payment—which may mean she does not buy. In the used car market, it is possible to overcome asymmetries of information without trust. Sellers can ameliorate the risk (and get what

[10] One might argue that, across the whole population, it is better if riders pay as otherwise no-one would become a taxi driver. But it would still be in any individual rider's financial interests not to pay: we have an n-person prisoner's dilemma, and a purely self-regarding person would never pay for taxi rides.

their car is worth) by offering devices such as warranties. Even with asymmetric information, *caveat emptor,* or 'buyer beware', is the norm. But the situation in finance is not a simple market for lemons. We can identify at least two differences.

The first difference relates to the types of product that are for sale. In the market for lemons there are high quality cars and low quality lemons; the low quality lemons are still cars that someone would want to buy, at the right price, and the risk is that of over-paying. But there is a third type of car, which does not enter the marketplace. We can distinguish a lemon from a death trap, a car that has a potentially fatal fault, which is known to the seller. Death traps are not a feature of the used car market, they go on the scrap market instead. But in the run-up to the crash, the financial equivalent of death traps were sold, products that were not fit for purpose, for example vehicles for retirement savings that were expected to be completely wiped out in thirteen years (as discussed by Peyton Young and Noe in this volume).

The market for lemons is a simplified depiction of the used car market, and neither the model nor the market itself is characterized by the existence of death traps. This may be partly due to the long arm of the law, the existence of warranties, and the fact that (in the UK) buyers are savvy enough to insist on seeing a recent MOT certificate. But, importantly, most people would agree that it is simply wrong to sell a car that is a death trap to an unknowing or unwitting buyer. It would not even cross their minds to try to sell their death trap as a working car. We can trust people not to sell cars that are death traps, in the strong as well as the weak sense. But, before the financial crisis, some bankers knowingly sold the financial equivalent of death traps. There was not even the basic level of trust that is assumed in the market for lemons model.

A second difference between finance and the market for lemons is the modes of product-assessment that are available. The market for lemons is really a story about matching products to buyers, ensuring that buyers have enough information to make a relatively precise valuation of the product. A warranty is a signal of a high quality product, which is worth paying more money for. But warranties do not exist in finance. There are no guarantees; products come with a health warning stating that they can go down as well as up. Potential buyers need reliable information about risk and returns in order to purchase appropriate products.

An alternative device for getting product information is third party assessment. This sometimes operates in the used car market. When people buy and sell cars through personal ads there is unlikely to be a warranty, nor are most buyers competent to assess a car themselves. They may engage the services of a mechanic, a third party with the expertise to assess the car. In the 'assessor model' the burden of trustworthiness is shifted to the assessor and the relationship

between buyer and seller is governed by caveat emptor. Another example of this model is the housing market. A potential buyer pays third parties, in the UK a surveyor and a solicitor, to get all the relevant information. Estate agents, who represent the seller, are generally considered untrustworthy.

We can distinguish assessment from advice. Mechanics and surveyors give information about the quality, possibly including a valuation, but they don't generally give advice about whether the car or house will meet the buyer's needs, and whether or not to purchase it. In contrast, we sometimes expect not only information but also advice. In medicine, patients rely on the advice of doctors to make an informed choice between treatments. Sometimes, the doctor makes the choice for the patient, for instance choosing which of a number of possible drugs to prescribe. In the legal profession, we may expect our lawyer both to represent us and to advise us on the best course of action. The relationship between advisor and advisee involves strong trust. An advisor is supposed to take the advisee's interests into account, providing advice about how best to further them. Professions that operate on the advisor model often have a professional ethic, an idea of the standards of good service and a commitment to uphold those standards.

However, advice and assessment are not completely distinct. Assessors do not necessarily advise, but advisors must to be able to make an assessment in order to give good advice.

The existence of competent third party assessors allows the relationship between buyer and seller to remain caveat emptor. But third-party assessment failed in the run-up to the financial crisis. External ratings agencies completely underestimated the risk of some products. There are a variety of reasons for this failure, but it seems that at least some financial products are not amenable to third party assessment. They are so complicated that only those who construct them (or not even those who construct them!) are in a position to know all their implications.

The situation in finance may involve 'asymmetric expertise'. The market for lemons model concentrates on asymmetric information, knowing 'that' a product is a lemon. In order to know that a product is a lemon (if there is no way of signalling it) we require the existence of an expert, who knows 'how' to assess the product.[11] If a product is so complicated that only the seller is in a position to provide a reliable assessment, then there is an asymmetry of expertise between the seller and everyone else, an asymmetry of knowing 'how' to make an assessment. On that case, only the seller is in a position to know relevant information about the product, so we need sellers to be trustworthy providers of information. (This may involve not just giving truthful information,

[11] For more on the relation between knowing that and knowing how, see Stanley (2011).

but also revealing all relevant information to a buyer who does not know what information she should ask for.)

Finance is not simply a market for lemons. In the run-up to the crisis some people in the sector lacked even the basic level of trustworthiness assumed by the lemons model; and yet strong trust is particularly important in finance because there may be asymmetric expertise as well as asymmetric information. Of course, one response to these problems is simply to regulate financial products—to ban death traps and to prevent products from becoming too complicated for third party assessment—in order to ensure that sellers are weakly trustworthy. However, regulation is only a partial solution. Regulations can be gamed and, whilst we can easily agree that some products are not fit for purpose, such as the savings vehicles discussed above, it will not always be so easy to distinguish what counts as a financial death trap as opposed to simply a very risky product that some informed consumer might buy; or to decide the trade-off between allowing the sale of a product that most people consider an unacceptable risk, even though the odd risk-seeker might choose to invest. Too much emphasis on regulation obscures a second way we can prevent untrustworthy behaviour, namely to increase strong trustworthiness.

6. Preventing Untrustworthy Behaviour

In order to design effective policies to prevent untrustworthy behaviour, we need to understand the causes of that behaviour.[12] The most salient cause of untrustworthy behaviour is the deliberate breaching of trust, where someone who knows that a trust has been placed in her purposefully breaches that trust for her own private gain. But that is not the only, or maybe even the most prevalent, cause of breaches of trust. We cannot always assume that everyone—the truster, the trustee, and any theorist or third party observer to the transaction—frames the situation the same way. If we recognize the role of 'framing' in decision-making, then we can

[12] What we should do to encourage trustworthy behaviour also depends on the motivation that underpins strongly trustworthy behaviour. For example, imagine a charity wanting to increase donations. If potential donors are motivated by esteem then the charity should offer to acknowledge the donations publicly, if donors are motivated by sympathy then the charity should make salient the plight of those who will be helped by the donation, and if donors are motivated by commitment then the charity should remind potential donors of the relevant normative imperative. But we can still come to some general conclusions and recommendations, even in the absence of a complete and accurate picture of what motivates human behaviour.

identify a second cause of breach of trust: when the truster and trustee frame the situation differently.

In our everyday lives, we negotiate many different relationships—our roles may include family member, friend, neighbour, worker, employer, client, consumer etc.—which are governed by different norms. These most obviously include norms of conduct, or behavioural norms, but there may also be norms about what emotions or motivations are appropriate, even if emotions and motivations are not something that we can always control. For instance, in market exchanges the range of motivations is not much restricted and there is licensed self-regard; at home the balance is tilted much more towards other-regarding-ness; and at work we may be committed to a professional ethic. It may be possible to abide by the behavioural norm without the expected motivation and emotion, but the lack of appropriate motivation and emotion is often seen as problematic (imagine parents who do not love their children or doctors who are not motivated by the well-being of their patients, even if they perform all the behaviours that we expect from people occupying those roles).

A pre-requisite for deciding to abide by the norms of a particular relationship is seeing that they apply. Hence the considerations that motivate a person depend on the relationship that she takes herself to be in, how she 'frames' her situation, and hence the norms that govern her interactions. A person can fail to be trustworthy because she does not frame an inter-action as one that requires strong trust, even though she would have been motivated to be trustworthy had she framed it differently. This contrasts with the paradigm of untrustworthy behaviour, which involves the recognition that a trust has been placed followed by a deliberate breach of that trust. As applied to finance, imagine a transaction where the buyer of a product believes the buyer-seller relationship operates on an advisor model, so she expects strong trustworthiness on the part of the seller. Contrast the situation where the seller knows how the buyer frames the transaction and sells her an unsuitable product anyway, in order to make a profit from a client, with the situation where the seller believes that the pair are merely buyer and seller, in a caveat emptor relationship, so it is the buyer's job to get her own assessment of the product.

It does not matter to the neoclassical agent how anyone frames a transaction because the principe of self-regard dictates that agents will always pursue their private advantage. However, most people are not 'knaves', to use Hume's term for people who are only ever motivated by their private interests (Hume, 1741). Nowadays, psychologists use 'psychopath' as a label for people who lack empathy and conscience, and measure these tendencies on a 'psychopathy scale' (Hare & Vertommen, 2003; Levenson, Kiehl, & Fitzpatrick, 1995). The idea is that psychopathy is a

personality trait, which people may have to varying degrees, and 'psychopaths' are people with pathologically high levels of the trait.[13] Only one percent of the population are psychopaths. The vast majority of people are capable of non-self-regarding motivations.

Non-psychopaths are are *trust responsive*, tending to fulfil trust when they believe that it has been placed in them (Bacharach, Guerra & Zizzo, 2007). The trust game probably underestimates people's capacity for trustworthiness because the laboratory is an artificial situation with no cues from real life and what constitutes appropriate behaviour is ambiguous. Researchers phrase their instructions in 'neutral' terms and normatively-laden labels are avoided. For example, subjects are referred to as 'actors' or 'participants', not 'trusters' and 'trustees'; the actions are called 'transferring money' rather than 'placing and fulfilling trust'. The typical subject is a student who has agreed to participate at least partly in order to make money. (In fact, in the trust game, CEOs are more trustworthy than students (Fehr & List, 2004).) The action of sending money is open to multiple interpretations: it could be seen as placing a trust, but it could also be seen as a gamble undertaken in the hope of making more money. Even if it is seen as placing a trust, the receiver may think that the interaction is not properly framed as a trust situation, that placing trust is not appropriate in the laboratory, and therefore not be motivated to respond in a trustworthy manner. These tendencies can be exacerbated by *motivated construal*, when an ambiguous situation is interpreted in a way that is consonant with a person's interests, without the person necessarily even being conscious of this.[14] The more ambiguous and atypical a situation is, the more we can expect there to be motivated construal.

The trustworthiness of non-psychopaths can be enhanced by making it clear that the situation is one of strong trust and, furthermore, that it is one where strong trust and trustworthiness are appropriate. However, there is a small percentage of psychopaths, who always act according to the principle of self-regard and who would deliberately breach trust

[13] Scored according to the Hare Psychopathy Checklist-Revised (PCL-R), a psychopath is someone who scores more than 30 out of a possible 40. The mean score in the general population is 2-4 out of 40, with more than half the population scoring zero or one and females scoring lower than males (Neumann, & Hare, 2008). Note that the PCL-R is not a psychiatric diagnosis—although we might expect that people who have high levels of psychopathy will also be diagnosed with 'antisocial personality disorder'.

[14] An infamous example is found in Hastorf and Cantril (1954), where fans of opposing teams in a dirty game each saw the other side as being the perpetrators of most fouls.

whenever it was to their advantage. These people will only ever be weakly trustworthy. Material incentives are needed to ensure their trustworthiness. As we might expect, psychopaths are over-represented in the prison population, a widely quoted estimate is 15-25% (Hare, 1996).[15]

Unfortunately, whilst psychopaths may need sanctions to elicit weakly trustworthy behaviour, sanctions can have a negative effect on strong trustworthiness. There is an inherent tension between strong trust and material incentives: if trusters use incentives then they are not placing strong trust. Imagine someone taking the role of truster in a trust game who says, 'I trust you to return money and I will sanction you if you don't'. The sentence makes sense if we parse it as 'I (weakly) trust you to return money because I will sanction you if you don't'. But it is strange to say 'I (strongly) trust you to return money and I will sanction you if you don't' because, if the person really (strongly) trusted the trustee, then she wouldn't need to threaten sanctions. The sanction is a signal that the relationship is not governed by strong trust. Sanctions substitute for strong trust.

Evidence from laboratory experiments on trust confirms this: using the threat of punishment in order to enforce high returns in the trust game backfires (Fehr & Rockenbach, 2003; Fehr & List, 2004). In a version of the trust game where trusters stated their desired back-transfer and were allowed to impose a fine if they received less than the desired amount, imposing a fine led to *less* money being sent back. (Note that the proceeds from the fines did not go to the truster, so there was no financial benefit to imposing fines other than any effect on the amount of money sent back.) The threat of sanctions also led to less money being returned when the sanctions were imposed by the experimenter, without the knowledge of the trusters (Houser, Xiao, McCabe, & Smith, 2008). This finding is consistent with a large literature which demonstrates that monetary rewards and punishments sometimes backfire—for instance, payments reduce work effort and the offer of monetary compensation reduces willingness to do civic duty—known as the 'motivation crowding effect' (Frey & Jegen, 2001).

[15] Hare's (1996) estimate is based on his experience of testing the US prison population. Obviously the figure depends on factors such as the rate of incarceration and how prisoners are split between gaols and mental institutions, but studies on other prison populations agree that the rate of psychopathy in prison is higher than in the general population, finding proportions of 3%-49% (Sullivan, E. A., & Kosson, 2006). The mean score amongst US prisoners is 22-24 (Hare, 1996), which is also substantially higher than the mean score of 2-4 in the general population (Neumann, & Hare, 2008).

In contrast, those subjects who forewent the punishment option in favour of strong trust found that their trust was rewarded (Fehr & Rockenbach, 2003; Fehr & List, 2004). In the trust game with the punishment option, if the fine was not imposed then the trustees sent more money back and trusters were better off—both compared to when the fine was imposed *and* compared to the version where there was no possibility of a fine. Not imposing a threat of punishment when one is on offer is a signal that the relationship involves strong trust.

However, the majority of subjects threatened sanctions, even though the threats decreased their expected earnings (Fehr & List, 2004). This may be related to an overly pessimistic view of other people's motivations: people believe that others are more motivated by extrinsic incentives than they are themselves (Heath, 1999). In fact, the overwhelming majority of people have motivational structures that include both self-regarding and non-self-regarding elements, and most people have a large capacity for non-self-regarding behaviour. Whether or not they behave in a non-self-regarding and, hence, strongly trustworthy manner depends the institutional structures and the types of relationships they perceive they are in.

When designing institutions, there is a tension between the need to threaten sanctions, in order to keep psychopaths in line, and the need to reinforce the perception of non-psychopaths that strong trust is appropriate. To some extent, the conflict in approaches is unavoidable. But it is not as bad as it looks, or as it is sometimes made out to be.

The tension between strong trust and the need for sanctions occurs when there is a mixture of psychopaths and non-psychopaths. But the composition of organizations is not fixed. We can increase strong trust in organizations by making them a less attractive place for psychopathic types. It is an open question whether there are more psychopaths in finance than in other sectors. Psychopathy is more prevalent amongst business leaders than in the population as a whole, with the proportion of psychopaths rising to four percent (Babiak & Hare, 2006; Babiak, Neumann, & Hare, 2010; Board & Fritzon, 2005). However, to my knowledge there is no study comparing business leaders to leaders in other fields and it seems likely that psychopaths, being ruthless and manipulative, will be over-represented in leadership positions in any field. But it is also plausible that professions offering high financial rewards are particularly attractive to psychopaths, since they do not get satisfaction from non-self-regarding aspects of a job. In general, we can affect the type of people that we attract into roles by varying the currency in which rewards are offered (Brennan & Hamlin, 2000). This is something to bear in mind when considering executive compensation, and other types of compensation that society can bestow such as honours and esteem.

Even within a sector that is composed of a cross-section of society, including some psychopaths, the tension between strong trust and sanctions is less than it seems from the experiments cited above. The literature on the counter-productive effect of sanctions focuses on financial penalties, abstracting from the social or moral sanctions which are often attached to punishment. For instance, in experimental trust games with sanctions, there is no mention of fines or any normative language, only 'conditional payoff cuts' (e.g. Houser, Xiao, McCabe & Smith, 2008). A payoff cut could be seen as a price, not as a punishment. Even fines operate in an ambiguous space between price and punishment. I certainly know people who regard risking a parking ticket as taking a gamble and the possibility of a fine as the price of parking. In a well-known experiment, fining parents who were late picking up their children up from nursery led to an increase in lateness; the parents treated the fine as a price that was worth paying (Gneezy & Rustichini, 2000). The idea that a punishment is a price is embodied in the economic analysis of crime. In those models, when a fine is threatened the cost of the prohibited activity is the amount a violator can expect to pay (the level of fine weighted by probability of getting caught), and agents treat this cost like the price of engaging in the activity. The natural extension of this approach, as noted by the economist Tyler Cowan (2013)—possibly tongue in cheek—would be to charge VAT on fines, as a consumption tax on the fined activity.

If agents are completely self-regarding, then thinking about sanctions in wholly material terms makes sense. But non-self-regarding agents will also be sensitive to social sanctions, like opprobrium or disapproval, and the moral sanctions of their own conscience. These social and moral aspects of sanctioning can be hugely important. Since punishment is usually accompanied by opprobrium and guilt, studying financial sanctions in isolation is distorting. In the absence of accompanying social sanctions, it may be natural to interpret financial sanctions as a price rather than a punishment, and prices elicit different behaviour to punishments. The law threatens punishment for murder and theft, but people do not usually argue that these laws increase the incidence of crime. In these cases it is clear that the behaviours are wrong, that there is a social concensus that they are inappropriate, and that the material sanctions are targeted at a small minority of deviants. Material punishments do not necessarily have counter-productive effects if they are embedded in a clear social and moral framework. If we have that framework then, in the same way that some have advocated using different types of rewards to motivate different types of people (e.g. Brennan & Hamlin, 2000), we can use different types of sanctions to motivate different types of people.

Of course, the reaction to the financial crisis has included moral opprobrium—from the general public. But general opprobrium alone may not be sufficient. People may have a generic dislike of criticism and censure but, in order to strongly motivate, opprobrium needs to come from those whose opinion they care about. The same point can be made about the motivating power of esteem: a complete theory will specify whose esteem is being pursued. Often the people whose approbation is sought is not that of society in general, but that of a smaller group of people who are close to or connected with the agent. For instance, in the professional domain, people may seek approbation from others within their organization or employment sector. These are also the people who set the norms of conduct within the workplace. So social sanctions that would effectively prevent untrustworthy behaviour in finance need to be imposed from within the sector and strongly trustworthy behavour needs to be supported by the culture of organizations.

These two approaches—changing the composition of organizations and changing their culture—are complementary. The entry and exit of those who are motivated by more than self-regard can be self-reinforcing (Bruni, & Smerilli, 2009). To the extent that the people at the top of organization are particularly influential in setting the tone, their behaviour and the sort of behaviour that they endorse can be disproportionately important. It is alleged that some leaders of financial institutions either deliberately breached the trust of their clients or tacitly endorsed untrustworthy behaviour of those further down the organization. It seems that some business leaders do need the to be motivated by sanctions. However, even amongst business leaders, the vast majority of people have a capacity for empathy and conscience. We shouldn't expect that they are all looking to take advantage of trusters. Regulation change is part of the response to the financial crisis, but culture-change is also a useful part of the policy toolbox.

Conclusion

When Mill introduced the 'arbitrary definition' that was to become the neoclassical economic agent, his idea was that the principle of self-regard was an approximation, which "is then to be corrected by making proper allowance for the effects of any impulses of a different description, which can be shown to interfere with the result in any particular case" (Mill, 1936). The principle of self-regard abstracts away from the non-self-regarding motivations that result in strong trustworthiness. Strong trust is ubiquitous, efficient, and has a priority over weak trust. If we want to encourage trustworthy behaviour in finance, then we should not discount non-self-regarding motivations, and we should design policies and institutions that encourage strong trustworthiness.

References

Abbink, K., & Sadrieh, A. (2009). The pleasure of being nasty. *Economics Letters*, *105*(3), 306-308.
Collard, D. (2001). Edgeworth's propositions on altruism. *Economic Journal*, *85*(338), 355-60.

Akerlof, G. A. (1970). The market for" lemons": Quality uncertainty and the market mechanism. *The quarterly journal of economics*, 488-500.

Babiak, P., & Hare, R. D. (2009). *Snakes in suits: When psychopaths go to work*. HarperCollins.

Babiak, P., Neumann, C. S., & Hare, R. D. (2010). Corporate psychopathy: Talking the walk. *Behavioral sciences & the law*, *28*(2), 174-193.

Bacharach, M. (1999). Interactive team reasoning: a contribution to the theory of cooperation. *Research in Economics 53*, 117–147.

Bacharach, M. (2006). *Beyond Individual Choice: Teams and Frames in Game Theory*. N. Gold and R. Sugden (eds). Princeton: Princeton University Press.

Bacharach, M., Guerra, G., & Zizzo, D. J. (2007). The self-fulfilling property of trust: An experimental study. *Theory and Decision*, *63*(4), 349-388.

Baier, A. (1986). Trust and antitrust. *Ethics*, *96*(2), 231-260.

Batson, C. D. (2011). *Altruism in humans*. Oxford University Press.

Batson, C. D., & Shaw, L. L. (1991). Evidence for altruism: Toward a pluralism of prosocial motives. *Psychological Inquiry*, *2*(2), 107-122.

Berg, J., Dickhaut, J., & McCabe, K. (1995). Trust, reciprocity, and social history. *Games and economic behavior*, *10*(1), 122-142.

Board, B. J., & Fritzon, K. (2005). Disordered personalities at work. Psychology, Crime & law,

11 (1), 17-32.

Brennan, G., & Hamlin, A. (2000). *Democratic devices and desires*. Cambridge University Press.

Brennan, G., & Pettit, P. (2004). The economy of esteem: An essay on civil and political society. Oxford University Press.

Brewer, M. B., & Gardner, W. L. (1996). Who is this "we"? Levels of collective identity and self representations. Journal of Personality and Social Psychology, 71, 83-93.

Brewer, M. B., & Kramer, R. M. (1986). Choice behavior in social dilemmas: Effects of social identity, group size, and decision framing. *Journal of personality and social psychology*, *50*(3), 543.

Broome, J. (1992). Deontology and economics. *Economics and Philosophy*, *8*(2), 269-282.

Brown, C. (2011). Consequentialize This. *Ethics*, *121*(4), 749-771.

Bruni, L., & Smerilli, A. (2009). The value of vocation. The crucial role of intrinsically motivated people in values-based organizations. Review of social economy, 67(3), 271-288.

Collard, D. (2001). Edgeworth's propositions on altruism. Economic Journal, 85(338), 355-60.

Collard, D. A. (1978). *Altruism and economy: A study in non-selfish economics*. Oxford: Martin Robertson.

Cowan, T. (2013). How you know that Singaporeans are really serious about microeconomics. Retrieved August 19, 2013, from: http://marginalrevolution.com/marginalrevolution/2013/08/how-you-know-that-singaporeans-are-really-serious-about-microeconomics.html

Deci, E. (1975) *Intrinsic Motivation*. New York: Plenum Press.

Edgeworth, F. Y. (1881). Mathematical Psychics: An Essay on the Application of Mathematics to the Moral Sciences. London: Kegan Paul.

Falk, A., Fehr, E., & Fischbacher, U. (2008). Testing theories of fairness—Intentions matter. *Games and Economic Behavior*, *62*(1), 287-303.

Fehr, E. (2009). On the economics and biology of trust. *Journal of the European Economic Association*, *7*(2-3), 235-266.

Fehr, E., & Fischbacher, U. (2003). The nature of human altruism. *Nature*, *425*(6960), 785-791.

Fehr, E., & List, J. A. (2004). The hidden costs and returns of incentives—trust and trustworthiness among CEOs. *Journal of the European Economic Association*, *2*(5), 743-771.

Fehr, E., & Rockenbach, B. (2003). Detrimental effects of sanctions on human altruism. *Nature*, *422*(6928), 137-140.

Frey, B. S., & Jegen, R. (2001). Motivation crowding theory. *Journal of economic surveys*, *15*(5), 589-611.

Gneezy, U., & Rustichini, A. (2000). Fine Is a Price, A. J. Legal Stud., 29, 1-17.

Gold, N. (2012). Team reasoning and cooperation. In S. Okasha and K. Binmore (eds). *Evolution and Rationality: Decisions, Cooperation and Strategic Behaviour.* Cambridge: Cambridge University Press

Gold, N. & Sugden, R. (2007). Theories of team agency. In F. Peter and S. Schmidt (eds). *Rationality and Commitment*. Oxford: Oxford University Press.

Hardin, R. (2004). Trust and trustworthiness (Vol. 4). Russell Sage Foundation.

Hardin, R. (1996). Trustworthiness. Ethics, 107(1), 26-42.

Hare, R. D. (1996). Psychopathy: A Clinical Construct Whose Time Has Come. *Criminal Justice and Behavior.* 23(1): 25-54.

Hare, R. D., & Vertommen, H. (2003). *The Hare psychopathy checklist-revised*. Multi-Health Systems, Incorporated.

Hastorf, A. H., & Cantril, H. (1954). They saw a game; a case study. *The Journal of Abnormal and Social Psychology*, *49*(1), 129

Hawley, K. (2012). Trust, Distrust and Commitment 1. *Noûs*.

Heath, C. (1999). On the social psychology of agency relationships: Lay theories of motivation overemphasize extrinsic incentives. *Organizational behavior and human decision processes*, *78*(1), 25-62.

Holton, R. (1994). Deciding to trust, coming to believe. *Australasian Journal of Philosophy*, *72*(1), 63-76.

Houser, D., Xiao, E., McCabe, K., & Smith, V. (2008). When punishment fails: Research on sanctions, intentions and non-cooperation. *Games and Economic Behavior*, *62*(2), 509-532.

Hume, D. (1741) Of the Independency of Parliament. In *Essays Moral, Political, and Literary*. Retrieved August 19, 2013, from: http://www.gutenberg.org/files/36120/36120-h/36120-h.htm

Jevons, W. S. (1871). *The theory of political economy*. Reprint, New York, D. Appleton and company, 1880. Retrieved August 19, 2013, from: *http://www.gutenberg.org/ebooks/33219*

Jones, K. (1996). Trust as an affective attitude. *Ethics*, *107*(1), 4-25.

Levenson, M. R., Kiehl, K. A., & Fitzpatrick, C. M. (1995). Assessing psychopathic attributes in a noninstitutionalized population. *Journal of personality and social psychology*, *68*(1), 151.

Lewin, S. B. (1996). Economics and psychology: Lessons for our own day from the early twentieth century. *Journal of Economic Literature*, *34*(3), 1293-1323.

Mackie, G. (unpublished paper) "Effective Rule of Law Requires Construction of A Social Norm of Legal Obedience"

Mandler, M. (2001). A difficult choice in preference theory: rationality implies completeness or transitivity but not both. *Varieties of Practical Reasoning*, 373-402.

Marshall, A. (1890) *Principles of Economic Thought.* Reprint at Rod Hay's Archive for the History of Economic Thought, McMaster University, Canada. Retrieved August 19, 2013, from: *http://socserv2.socsci.mcmaster.ca/~econ/ugcm/3ll3/index.html*

Mas-Colell, A., Whinston, M. D., & Green, J. R. (1995). *Microeconomic theory* (Vol. 1). New York: Oxford university press.

Mill, J. S. (1836). 'On the"Definition of Political Economy; and on the Method of Investigation Proper to It' *London and Westminster Review*, October 1836. Reprint, Mill, *Essays on Some Unsettled Questions of Political Economy*. 1844. Reprint Project Gutenberg 2004. Retrieved August 19, 2013, from: http://www.gutenberg.org/ebooks/12004

Mill, J. S. (1859). *On Liberty.* Reprint, Mill, J. S. (1909). *Harvard Classics: Volume 25*. Collier. Retrieved August 19, 2013, from: http://ebooks.adelaide.edu.au/m/mill/john_stuart/m645o/

Neumann, C. S., & Hare, R. D. (2008). Psychopathic traits in a large community sample: links to violence, alcohol use, and intelligence. *Journal of Consulting and Clinical Psychology*, *76*(5), 893.

Offer, A. (1997). Between the gift and the market: the economy of regard. *The Economic history review*, *50*(3), 450-476.

Rabin, M. (1993). Incorporating fairness into game theory and economics. *The American economic review*, 1281-1302.

Sen, A. K. (1977). Rational fools: A critique of the behavioral foundations of economic theory. *Philosophy & Public Affairs*, *6*(4), 317-344.

Smith, A. (1759) *The Theory of Moral Sentiments.* Reprinted, Metalibri 2005. Retrieved August 19, 2013, from: http://metalibri.wikidot.com/title:theory-of-moral-sentiments:smith-a

Smith, A. (1776) *An Inquiry into the Nature and Causes of the Wealth of Nations*. Retrieved August 19, 2013, from: http://www.gutenberg.org/ebooks/3300

Smith, V. (2013). 'Adam Smith: From Propriety and Sentiments to Property and Wealth' *Forum for Social Economics.*

Sober, E., & Wilson, D. S. (1998). *Unto others: The evolution and psychology of unselfish behavior* (No. 218). Harvard University Press.

Stanley, J. (2011) *Know How*, Oxford University Press.

Sullivan, E. A., & Kosson, D. S. (2006). Ethnic and cultural variations in psychopathy. *Handbook of psychopathy*, 437-458.

Sugden, R. (1993). 'Thinking as a team: toward an explanation of nonselfish behavior'. Social Philosophy and Policy, vol 10, pp. 69-89.

Zizzo, D. J., & Oswald, A. J. (2001). Are People Willing to Pay to Reduce Others' Incomes?. *Annales d'Economie et de Statistique*, 39-65.